Compression of Deep Neural Networks for Efficient Real-Time Data Analytics in the Resource-Constrained Internet of Things

PhD Synopsis

Name of Scholar	**Tawseef Ahmad Dar**
Enrolment Number	21165CUKmr002
Date of Registration	20 September 2021
Programme	Ph.D Computer Science & Engineering
Supervisor	**Dr. Mohammad Ahsan Chishti** *Associate Professor & Dean* School of Engineering & Technology Central University of Kashmir
Department	**Information Technology**
School	**School of Engineering and Technology**

Central University of Kashmir
Ganderbal, J&K

AUGUST 2022

Compression of Deep Neural Networks for Efficient Real-Time Data Analytics in the Resource-Constrained Internet of Things

1) Introduction: Internet of Things (IoT) is set to revolutionize all aspects of our lives. IoT ecosystems are composed of smart objects, i.e., objects with sensing and processing capabilities that can understand and react to their environment using embedded sensors [1]. The combination of various emerging technologies such as near-field communications, real-time localization, embedded sensors, and the internet enable us transform everyday objects into smart objects. IoT offers a platform for sensors and devices to communicate seamlessly within a smart environment and enables information sharing across platforms in a convenient manner. Over 30 billion [2] devices ranging from smart phones, laptops, sensors, and game consoles are anticipated to be connected to the Internet through several heterogeneous access networks enabled by technologies, such as radio frequency identification (RFID) and wireless sensor networks.

The continuum of devices and objects are interconnected through a variety of communication solutions, such as Bluetooth, WiFi, ZigBee, and GSM. The volume of data generated by sensors, devices, social media, health care applications and various other software applications and digital devices that continuously generate large amounts of structured, unstructured, or semi-structured data is strongly increasing. This massive data generation results in Big data [3]. The term Big data refers to extremely large, heterogeneous and complex data-sets, which cannot be handled by the conventional data processing and storage applications such as Relational Database Management System (RDBMS) [4]. In contrast to the traditional data, big data mainly differs in the following way [5]: (i) data rate is more rapid and data volume is constantly updated, (ii) data is of semi-structured or unstructured nature, (iii) data sources are fully distributed, (iv) data access is in batch mode

or real time instead of more interactive feature in the traditional data, and (v) integration of heterogeneous data from different sources becomes complicated.

Data Analytics refers to the analysis of every fragment of data generated from IoT devices at right time in order to extract intelligent insights [6]. Data analytics means processing every segment of data to discover the patterns in the data, dig out the hidden knowledge, extract precious information. The key concern of IoT data analytic systems is to attain enhanced elucidation of data so that efficient control decisions can be devised [7]. Machine Learning (ML) would undoubtedly play a role in generating valuable inferences from this massive volume of data and hence will assist in creating smarter IoT. In this regard, exploring the potential of ML for IoT data analytics becomes highly crucial.

2) Review of Literature and Identification of Research Gaps

2.1) Literature Review: In IoT environments, big data may be generated from a variety of application scenarios ranging from smart home, smart city, smart transportation to smart healthcare applications. The heterogeneous data generated from IoT devices may have certain statistical and strong correlative features across several dimensions such as time and location, and also the devices may have social relations among themselves [8]. The data analytics life-cycle to handle IoT data starts with the collection of the raw data and then the acquired raw information is aggregated at the aggregator/gateway, and is subsequently sent to the cloud for further processing [9]. The raw information gathered from the IoT sensors may represent various parameters such as vibration, pressure, temperature, motion, heart rate etc., and they need to be converted into recognizable formats for further analysis. The acquired data may be transmitted to the cloud centre either in a single hop or multiple hops based on the employed network topology. The completion of this life-cycle requires effective coordination of various activities among IoT sensors, IoT aggregator nodes/gateways, in-transit network devices such

as routers and software/hardware resources in the cloud centre [10]. The key technology enablers for IoT data analytics depicted in Fig 1 are:

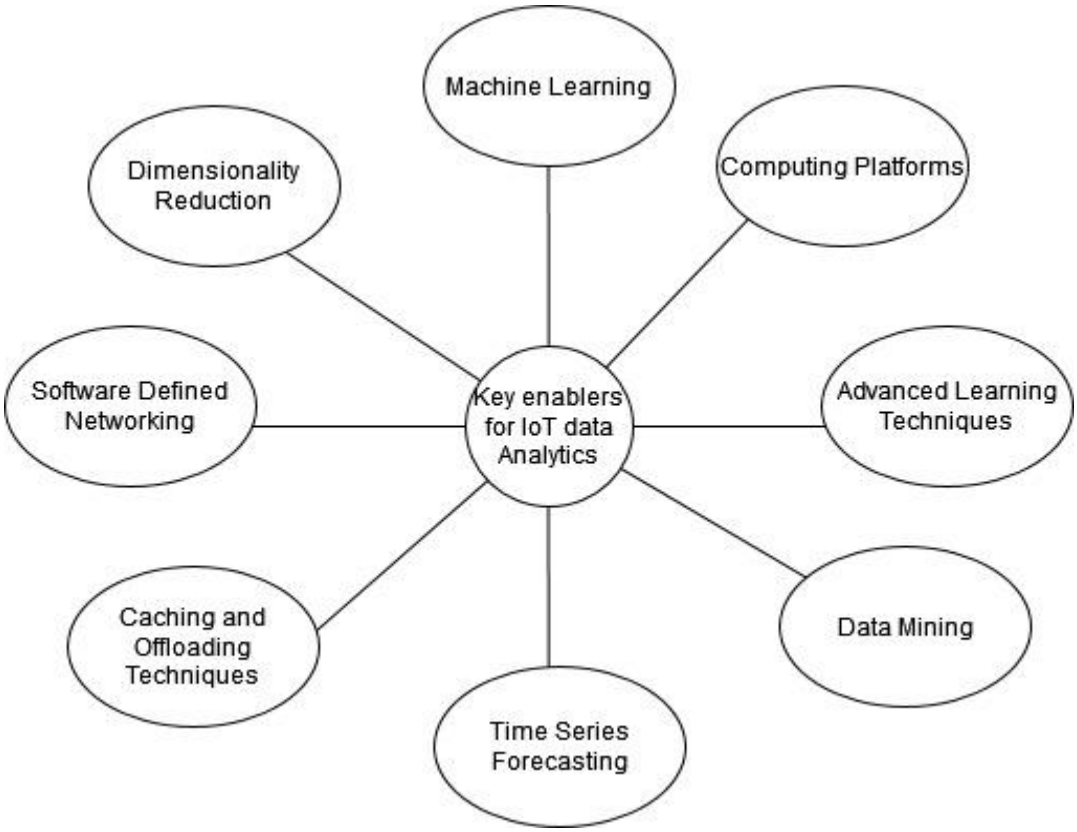


Fig 1: Key enablers for IoT data analytics.

2.1.1) Data mining: Data mining is used to extract implicit information from the datasets and transform this information to a known structure for further usage by employing suitable methods like classification, clustering, prediction, association rule mining. Data analytics aim to extract knowledgeable information that helps in making predictions, identifying recent trends, finding hidden information, and ultimately, making decisions [13].

2.1.2) Machine learning: Machine learning is an application of artificial intelligence that provides ability to systems to automatically learn and improve from experience without being explicitly programmed. These techniques aim to create a functional

relationship between input data-set and output actions, and are capable of performing predictions and decisions based on the input data without requiring the need of following static program instructions [14]. These techniques are broadly categorized into unsupervised, supervised and reinforcement learning.

2.1.3) Advanced learning techniques: In addition to the conventional machine learning techniques, several advanced learning techniques such as deep learning, transfer learning, and incremental learning can be utilized to extract useful information from complex or incomplete data-sets. Deep learning is suitable for modelling complex behaviours of heterogeneous data-sets while transfer learning is mostly useful for scenarios with limited datasets .On the other hand, incremental learning deals with learning in real-time and is useful for applications where data arrives in a sequential order [14].

2.1.4) Time Series Forecasting: Among all types of IoT data, time series data (e.g., data from sensors indexed in time order) are becoming more and more widespread. They need to be stored and analysed. Time series data are time-dependent and each variable from a time series sequence is correlated with the next one. This is called as autocorrelation. Time series data are mostly used to predict / forecast the future instances of a measured variable based on its past observations [15]. Time series modelling techniques include Auto Regressive Integrated Moving Average (ARIMA), Hidden Markov Model (HMM) and Recurrent Neural Networks (RNN).

2.1.5) Computing platforms: With the rapid development of Internet of Things, the traditional centralized cloud computing is encountering severe challenges, such as high latency, low spectral efficiency, and non-adaptive machine type of communication. Motivated to solve these challenges, new technology is driving a

trend which shifts the function of centralized cloud computing to the edge devices of the network [14].

2.1.6) Dimensionality Reduction: Data pre-processing is an important part for effective machine learning and data mining. Most machine learning and data mining techniques may not be effective for high dimensional data. Dimensionality simply means the number of features (i.e. input variables) in the dataset. When the number of features is very large relative to the number of observations in the dataset, certain algorithms struggle to train effective models. This phenomenon is called as the Curse of Dimensionality [16]. To combat this curse of Dimensionality phenomenon, data downsizing techniques have been designed. These techniques are broadly categorised into two types: Feature Selection and Feature Extraction. Feature selection techniques filter irrelevant or redundant features from the dataset. The key difference between feature selection and extraction techniques is that feature selection techniques keep a subset of the original features while feature extraction techniques create new features from the existing ones [17]. By reducing the dimension of data, a significant gain can be achieved in saving the cost for storage, processing and communication resources.

2.1.7) Caching and offloading techniques: Caching techniques are required for efficient utilization of front haul/backhaul bandwidth and for maximization of content delivery efficiency. And data offloading techniques decide which data to offload to the cloud and which to process at the edge-side [14].

2.1.8) Software Defined Networking (SDN): Software-defined networking (SDN) technology is a novel approach to cloud computing that facilitates network management and enables programmatically efficient network configuration in

order to improve network performance and monitoring. SDN is meant to address the fact that the static architecture of traditional networks is decentralized and complex while current networks require more flexibility and easy troubleshooting. SDN suggests to centralize network intelligence in one network component by disassociating the forwarding process of network packets (Data Plane) from the routing process (Control plane) [18].

2.2) Research Gap

The enormous amount of valuable data generated from IoT environs need to be examined so as to reveal trends, unseen patterns, hidden correlations and actionable insights. The key challenge here is how to dig out useful information from such voluminous amounts of data, as such data are not useful without analytic power. Numerous solutions have been formulated to obtain valuable insights. However, these solutions are still in their developing stages. Cloud computing is considered as a useful paradigm for handling such massive amount of data by providing centralized pool of storage. However, it is not suitable for scenarios demanding low latency, real time operation and high spectral efficiency. Such applications include autonomous driving, posture recognition, fire detection, security monitoring etc. Hence, new technology shifts the functioning of centralized cloud to the edge devices of the network. This new paradigm is known as Fog computing. This would decrease network congestion as well as accelerate analysis and resulting decision making. The best possible solution for time stringent applications is to bring streaming data analytics closer to IoT data source. However, bringing analytics closer to IoT data source has its own set of challenges like limitation of computing, energy and storage resources. The main requirement of most of the IoT applications is an intelligent learning mechanism which can perform tasks like prediction, classification or clustering. However traditional learning approaches do not address the emerging analytical demands of IoT systems. Deep learning is

a promising approach to achieve the desired analytics in IoT applications. Deep learning models are better than traditional machine learning approaches in two ways. First they reduce the need for engineered feature sets to be used for training. Therefore, features that might not be perceptible to a human can be extracted easily by Deep Learning models. Furthermore, Deep Learning models improve the accuracy of prediction models. Deep learning (DL) has shown remarkable results in image recognition, speech recognition, physiological and psychological state detection etc. and these services form the foundational services for IoT applications. Also, these services need fast data analytics in order to be efficient in their context. Hence, adapting deep learning models for deployment on resource constrained devices becomes a crucial point of study. A recently explored solution space lies in compressing the deep neural network before using on the device. With the aid of compression, a dense network is converted to a sparse network by exploring the redundancy in model without hurting the model accuracy. Although this approach seems promising for getting close to on node analytics, more investigations need to be performed as it is still in the early stages of development.

3) Objectives and Methodology

3.1) Objectives: The main objective of our research is to study network compression and approximate computing approaches for Deep Neural Networks in order to ensure deployment of Neural Networks (Deep Learning models) on resource constrained IoT devices. In order to achieve our research goal, following objectives need to be attained:

i. To investigate the applicability of network compression and approximate computing approaches to a variety of DL architectures.
ii. To carry out the proficiency analysis of the compressed deep neural network models.
iii. To specify the trade-off between the rate of compression and accuracy of a deep neural network, as more compression leads to degraded accuracy.

3.2) Methodology: Given the importance of deployment of deep neural networks on resource constrained IoT devices, the study would seek to understand the topic in detail. A careful examination of the work of researchers would help identify further shortcomings in the existing systems. To achieve the research objectives different data sets will be collected on which the different deep neural networks has been trained. Then after the collection of data sets compression techniques will be incorporated in these deep neural network models. After that accuracy of the compressed models will be compared with uncompressed models. Then analyse how a deep neural network's accuracy and compression rate are traded off. A solution to address the identified loopholes would be designed and formulated into a mathematical model. Necessary simulation would be carried to compare it with known methods.

References

[1] Behailu Negash, Amir M. Rahmani, Pasi Liljeberg, and Axel Jantsch, "Fog Computing Fundamentals in the InternetofThings,"Springer10.1007/978-3-319-57639-8_1.

[2] https://spectrum.ieee.org/tech-talk/telecom/internet/popular-internet-of-things-forecast-of-50-billiondevices-by-2020-is-outdated[Published:April16,2017].

[3] K.Kambatla,''Trends in big data analytics,''J.Parallel Distrib.Comput., vol. 74, no. 7, pp. 2561–2573, 2014.

[4] S. Yin and O. Kaynak, ''Big data for modern industry: Challenges and trends [point of view],'' Proc. IEEE, vol.103,no.2,pp.143–146,Feb.2015.

[5] H. Hu, Y. Wen, T.-S. Chua, and X. Li, ''Towards scalable systems for big data analytics: A technology tutorial,'' IEEE Access, vol. 2, pp. 652–687, Jul. 2014.

[6] Shikhar Verma, Yuichi Kawamoto, Zubair Md. Fadlullah, Hiroki Nishiyama, and Nei Kato, A Survey on Network Methodologies for Real-Time Analytics of Massive IoT Data and Open Research Issues, IEEE Communications Surveys and Tutorials, Vol. 19, No. 3, pp. 1457-1477, 2016.

[7] K. Panetta, Gartner's top 10 strategic technology trends for 2017, Smarter with Gartner 18.

[8] S.Bi,R.Zhang,Z.Ding,andS.Cui,''Wireless communications in the era of big data,'' IEEE Commun.Mag.,vol.53,no.10,pp.190–199,Oct. 2015.

[9] H. Cai, B. Xu, L. Jiang, and A. V. Vasilakos, ''IoT-based big data storage systems in cloud computing:

Perspectivesandchallenges,''IEEEInternetThingsJ.,vol.4,no.1,pp.75–87,Feb.2017.

[10] D.Puthal,S.Nepal,R.Ranjan,andJ.Chen,''Threats to networking cloud and edge data centers in the Internet ofThings,''IEEECloudComput.,vol.3,no.3,pp.64–71,May2016.

[11] Chun-Wei Tsai, Chin-Feng Lai, Ming-Chao Chiang, and Laurence T. Yang, "Data Mining for Internet of Things: A Survey," IEEE Communications Surveys & Tutorials, VOL. 16, NO. 1, FIRST QUARTER 2014.

[12]Mohsen Marjani et al., "Big IoT Data Analytics: Architecture, Opportunities, and Open Research Challenges,"DigitalObjectIdentifier10.1109/ACCESS.2017.2689040.

[13]C.-W. Tsai, ''Big data analytics: A survey,'' J. Big Data, vol. 2, no. 1, pp. 1–32, 2015.

[14]Shree Krishna Sharma, and Xianbin Wang, "Live Data Analytics With Collaborative Edge and Cloud Processing in Wireless IoT Networks," IEEE Access 10.1109/ACCESS.2017.2682640.

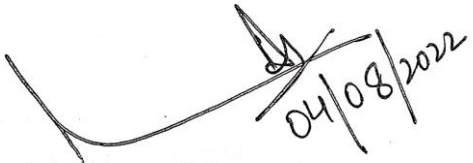
[15]http://iot-ee.com/en/2017/08/07/analysing-iot-data-introduction-time-series-forecasting-python/ [Published: August7,2017].

[16]http://www.cs.binghamton.edu/~lyu/SDM07/DR-SDM07.pdf

[17]https://elitedatascience.com/dimensionality-reduction-algorithms [Published: May 23,2017]

[18]D. Kreutz, F. Ramos, P. E. Veríssimo, C. E. Rothenberg, S. Azodolmolky, and S. Uhlig, ''Software-defined networking: A comprehensive survey,'' Proc. IEEE, vol. 103, no. 1, pp. 14–76, Jan. 2015.

Signature of Research Scholar



Signature of Supervisor